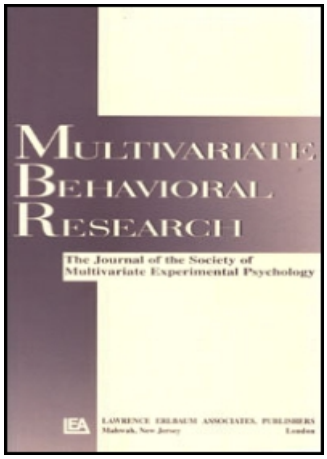
Testing hypotheses about the person-response function in person-fit analysis

Emons, W.H.M.; Sijtsma, K.; Meijer, R.R.

Published in:
Multivariate Behavioral Research

Publication date:
2004

Document Version
Publisher's PDF, also known as Version of record



Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

Testing Hypotheses About the Person-Response Function in Person-Fit Analysis

Wilco H. M. Emons [a]; Klaas Sijtsma [a]; Rob R. Meijer [b]
[a] Tilburg University.
[b] University of Twente.

Online Publication Date: 01 January 2004
To cite this Article: Emons, Wilco H. M., Sijtsma, Klaas and Meijer, Rob R. (2004) 'Testing Hypotheses About the Person-Response Function in Person-Fit Analysis', Multivariate Behavioral Research, 39:1, 1 - 35
To link to this article: DOI: 10.1207/s15327906mbr3901_1
URL: http://dx.doi.org/10.1207/s15327906mbr3901_1

Multivariate Behavioral Research, *39* (1), 1-35

Testing Hypotheses About the Person-Response Function in Person-Fit Analysis

Wilco H. M. Emons and Klaas Sijtsma
Tilburg University

Rob R. Meijer
University of Twente

The person-response function (PRF) relates the probability of an individual's correct answer to the difficulty of items measuring the same latent trait. Local deviations of the observed PRF from the expected PRF indicate person misfit. We discuss two new approaches to investigate person fit. The first approach uses kernel smoothing to estimate continuous PRF estimates. Graphical displays of PRFs were used to localize and diagnose misfit. The second approach approximates the PRF by a logistic regression model. Hypothesis tests on the regression parameters were used to detect certain types of misfit. A simulation study was conducted to investigate the Type I error rates and the detection rates of the regression approach.

A test score may be affected by factors other than ability. For example, a respondent may copy the correct answers to relatively difficult items in an exam from a high-ability neighbor (Cizek, 1999), suffer from test anxiety (Haladyna, 1994), fumble on the first items of the test (Meijer, 1994a), use an idiosyncratic interpretation of item content due to language difficulties (Van der Flier, 1980), or lack particular knowledge not covered by the school curriculum but required by the test (Harnisch & Linn, 1981). The result is an item-score vector that cannot be fitted by a hypothesized item response theory (IRT) model or that deviates from the observed item-score vectors of the majority of the test takers in the sample. The purpose of person-fit research is to identify such misfitting item-score vectors. The methods proposed here are general in the sense that they may be used to detect different kinds of aberrant response behavior including those mentioned in the examples.

Several authors discussed the person-response function (PRF) in the context of person-fit analysis (Lumsden, 1978; Nering & Meijer, 1998;

Correspondence concerning this article should be addressed to Wilco H.M. Emons, Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands.

Reise, 2000; Sijtsma & Meijer, 2001; Strandmark & Linn, 1987; Trabin & Weiss, 1983). Sijtsma and Meijer (2001) used the PRF as a tool to investigate person fit in nonparametric item response theory (NIRT; e.g., Sijtsma & Molenaar, 2002, p. 94) models. At the individual level, the PRF, to be defined in greater detail below, relates the probability of a correct answer to an item-difficulty scale. Local deviations of the observed PRF from the expected PRF under an IRT or NIRT model indicate misfit. Unlike person-fit statistics, such as statistic l_z (Drasgow, Levine, & Williams, 1985), that are used to classify *complete* item-score vectors as fitting or misfitting, the PRF can be used to identify *subsets* of item scores that disagree with the expected item scores. This may help the researcher to explain the observed misfit. For example, a PRF may show relatively high probabilities of a correct response to difficult items whereas success probabilities for easier items were lower. This may suggest that the respondent copied the answers to the difficult items from a high-ability neighbor at the exam. Thus the PRF enables a diagnostic approach to person-fit analysis (Reise, 2000).

In this study, we discuss two new approaches that use the PRF to investigate person fit under NIRT models. The first approach uses kernel smoothing (e.g., Simonoff, 1996) to obtain a (quasi-) continuous estimate of the PRF. Graphical displays of such smooth estimated PRFs can be used to localize the misfit and suggest specific types of aberrant response behavior. In the second approach, the PRF is modeled by a logistic regression model (e.g., Agresti, 1990; Fox, 1997). Hypothesis tests on the parameters of the logistic regression model are used to investigate certain types of person misfit. Using logistic regression for person-fit assessment was pursued earlier by Strandmark and Linn (1987) in the context of parametric IRT (e.g., Hambleton & Swaminathan, 1985) and by Reise (2000) in a multilevel modeling context.

The PRF approaches using kernel smoothing and logistic regression have properties that make them attractive for person-fit analysis. First, in practice person-fit analysis is based on a limited number of items, typically ranging from 20 to 100. Both graphical analysis using kernel smoothing and logistic regression are suited for analyzing such small data sets. Second, in contrast to the PRF approaches discussed by Trabin and Weiss (1983), Nering and Meijer (1998), and Sijtsma and Meijer (2001), kernel smoothing and logistic regression do not require a division of the item-score vector into disjoint item subsets. For example, Sijtsma and Meijer (2001) divide a 40-item test into five 8-item subsets of increasing difficulty, and compare the item scores between subsets. An outcome may be that performance on the fourth subset is better than that on the easier second and third subsets. This may be taken as evidence of misfit. One possible drawback of this

comparison is that the amount of information is reduced from 40 data points to five mean item scores, making a functional analysis less effective. Another drawback is that arbitrary decisions have to be made about the number and the size of these subsets, and that different decisions may lead to different conclusions. Third, the graphical approach is easy to implement and the resulting graphs of the PRFs are easy to understand. Therefore, the graphical method is suited for person-fit analysis by test practitioners in small scale settings, such as the classroom and individual psychological diagnostics. Fourth, logistic regression can be used to test the fit of item-score vectors against specific alternatives. Such directed tests may increase the power to detect specific types of aberrant response behavior (Meijer, 2003; Meijer & Sijtsma, 2001).

This article is organized as follows. First, we explain NIRT and give a formal definition of the PRF. Second, we explain kernel smoothing for obtaining (quasi-) continuous PRF estimates. Third, we discuss logistic regression models and their application to person-fit assessment. Fourth, we present the results of a simulation study that explored the usefulness of logistic regression in identifying person misfit.

Nonparametric Item Response Theory

We assume that the test consists of J items. Let X_j ($j = 1, ..., J$) be the binary random variable for item scores, with a value of 1 for a correct or a coded response, and a value of 0 otherwise. Furthermore, let $\mathbf{X} = (X_1, ..., X_J)$ be the random vector of item-score variables. Let $X_+ = \Sigma_j X_j$ denote the test score. Let π_j ($j = 1, ..., J$) denote the population proportion of persons with $X_j = 1$; and let $\hat{\pi}_j$ be the sample estimate of π_j. We assume that the J items are ordered and numbered such that $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_J$; that is, the items are ordered from easy to difficult and numbered accordingly, and subscript j denotes the rank number of the ordered item-score vector.

IRT models relate the probability of a correct answer to the latent trait θ by means of the item response function (IRF): $P_j(\theta) = P(X_j = 1|\theta)$. In NIRT, the IRFs are defined by order restrictions on the $P_j(\theta)$s (Sijtsma, 1998; Sijtsma & Molenaar, 2002, p. 14), but NIRT models refrain from a parametric definition of the IRF by some suitable parametric function, such as the logistic or the normal ogive. Compared with parametric IRT models, NIRT models have greater flexibility for fitting test data than their parametric counterparts (Junker & Sijtsma, 2001; Ramsay, 1991; Sijtsma & Molenaar, 2002, pp. 6, 15-16).

In this article, we use NIRT models that are based on the common assumptions of unidimensionality (UD), local independence (LI),

monotonicity (M), and the more restrictive assumption of nonintersecting IRFs. UD means that the latent trait that explains the examinee's performance is unidimensional; that is, θ is a scalar. LI means that the item responses are statistically independent conditional on θ; that is, $P(\mathbf{X} = \mathbf{x}|\theta) = \Pi_j P(X_j = x_j|\theta)$. Assumption M specifies that the IRFs are monotonely nondecreasing in θ; that is, for two arbitrarily chosen fixed values θ_a and θ_b, and assuming that $\theta_a < \theta_b$, we have that $P(\theta_a) \leq P(\theta_b)$. IRT models satisfying the assumptions of UD, LI, and M imply a stochastic ordering of θ by means of the test score X_+ (Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997). This property justifies using number-correct scores to order the respondents according to θ.

Finally, the assumption of nonintersecting IRFs states that for two items i and j and an arbitrary fixed value θ_0, if we know that $P_i(\theta_0) > P_j(\theta_0)$, then $P_i(\theta) \geq P_j(\theta)$, for all θ. This is the assumption of an invariant item ordering (IIO; Sijtsma & Junker, 1996). IRT models having an IIO imply the same item difficulty ordering for each θ (except for possible ties) and consequently, each subgroup from the population of interest. An IIO facilitates interpretation of test performance, for example, in intelligence testing, testing for developmental progress reflected in certain behavior sequences, differential item functioning, and person-fit analysis (Sijtsma & Junker, 1996). IIO can be evaluated in real test data using methods discussed by Sijtsma and Molenaar (2002; also, see Mokken & Lewis, 1982; Rosenbaum, 1987a, 1987b). The IIO assumption was fitted to data from, for example, an inductive reasoning test (De Koning, Sijtsma, & Hamers, 2002), a transitive reasoning test (Sijtsma & Junker, 1996; Verweij, Sijtsma, & Koops, 1996), a nonverbal abstract reasoning test (Meijer, 2003), and a child intelligence test (Emons, Sijtsma, & Meijer, 2002). Sijtsma and Molenaar (2002) provide several other examples of test data that were fitted by the IIO assumption.

An NIRT model that satisfies all four assumptions is Mokken's (1971) model of double monotonicity. This model may be seen as a nonparametric version of the Rasch (1960) model (see De Koning et al., 2002). Sijtsma and Meijer (2001) showed that all IRT models satisfying the assumptions of UD, LI, M, and IIO provide useful PRF definitions. They also showed that such person-fit methods based on a PRF definition are robust against mild violations of IIO (see also Emons, 2003).

The Person-Response Function

The PRF (Lumsden, 1978; Sijtsma & Meijer, 2001; Trabin & Weiss, 1983) describes for a person υ with latent trait value θ_υ the probability of a correct answer to items measuring θ as a function of their difficulties. Let S_υ be the random response variable for person υ and let $S_\upsilon = 1$ stand for a correct answer and $S_\upsilon = 0$ for an incorrect answer. We assume a continuous latent difficulty scale, denoted by δ, with δ_j being the location of item j. For fixed θ_υ, the PRF is defined by

$$P_\upsilon(\delta) \equiv P(S_\upsilon = 1|\delta). \tag{1}$$

Thus, analogous to the IRF, which describes the probability of a correct answer as a function of θ and fixed item parameters, the PRF describes the probability of a correct answer as a function of item difficulty and a fixed person parameter.

Sijtsma and Meijer (2001) defined the PRF in the context of NIRT. Let $G(\theta)$ be the cumulative distribution of θ, then the population item difficulty is defined as

$$1 - \pi_j = \int_\theta [1 - P_j(\theta)]dG(\theta),$$

which has domain [0, 1]. Substituting $1 - \pi_j$ for the item difficulty δ in Equation 1 and dropping index j gives the PRF,

$$P_\upsilon(1 - \pi) \equiv P(S_\upsilon = 1|1 - \pi).$$

Sijtsma and Meijer (2001) showed that $P_\upsilon(1 - \pi)$ is a nonincreasing function under NIRT models satisfying UD, LI, M, and IIO. Local deviations from this nonincreasingness can be used to identify types of misfit, and support the interpretation of possible causes of misfit. Figure 1 shows an example of three PRFs (solid lines) expected under the model of double monotonicity and one PRF (dashed line) indicating misfit at the relatively difficult items. Before we discuss methods to investigate fit of the PRF, we illustrate how the PRF is sensitive to different types of aberrant response behavior.

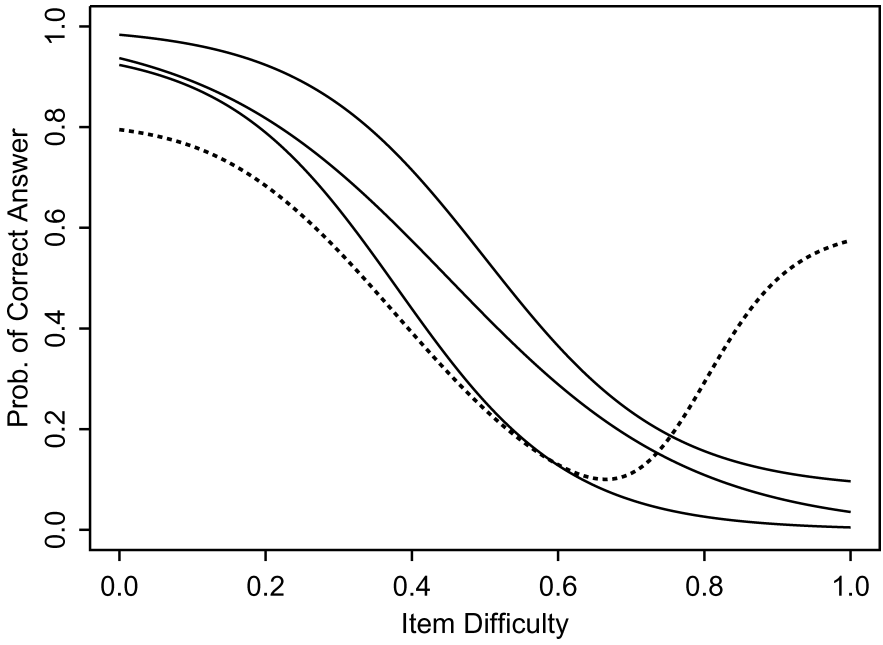


Figure 1
Example of Three Fitting PRFs (Solid Lines) and One Misfitting PRF (Dashed Line)

Person-Response Functions for Spuriously High Scores and Spuriously Low Scores

Aberrant response behavior, such as cheating, test anxiety, item disclosure, and misunderstanding of instructions, may have a detrimental effect on the validity of individual test scores (Haladyna, 1994; Meijer, 1994a, 1997). See Meijer (2003) for a discussion of person-fit statistics sensitive to different kinds of aberrant response behavior (e.g., Drasgow et al., 1985; Klauer, 1991; Meijer, 1994b). In this study, we distinguished two general classes of aberrant response behavior. The first class contains types of aberrant response behavior yielding unexpectedly many correct answers to relatively difficult items. This results in spuriously high X_+ scores. The second class contains types of aberrant response behavior that yield unexpectedly many incorrect answers to relatively easy items. This results in spuriously low X_+ scores. To identify both types of aberrant response behavior we used two prototypical PRFs.

Spuriously High X_+ Scores. Suppose a respondent of average ability took a 40-item exam and copied the correct answers to the 10 most difficult items from his/her high-ability neighbor. The resulting PRF for this respondent decreases for the first 30 items, and then increases for the 10 most difficult items (Figure 2a). In general, a PRF for response behavior causing misfit at the most difficult items is characterized by a U-shape.

Spuriously Low X_+ Scores. Suppose an average respondent suffering from severe test anxiety took a high-stakes test consisting of 40 items and gave incorrect answers to the first 10 items of the test, which were the easiest items. The result is a PRF that first increases and then decreases (Figure 2b). In general, this type of response behavior is characterized by a bell-shaped PRF.

Particular types of aberrant response behavior are characterized by a near horizontal PRF (Figure 2c). This happens, for example, when a respondent took a test with little preparation, mainly to familiarize with content and test format so as to increase the probability of passing a future test, but no serious intention to pass now. As a result, he/she guessed for the correct answers (e.g., Van den Brink, 1977). The interpretation of near horizontal PRFs is not straightforward because this shape is also expected for low-ability respondents. Both for scouting-respondents and low-ability respondents, the observed PRF agrees with the expected nonincreasing PRF. To investigate person fit based on horizontal PRFs, collateral information is needed, such as a respondent's test score on comparable tests and teacher's observations. Based on this information, the researcher may decide whether the observed horizontal PRF is suspicious and supplementary person-fit research is needed.

Graphical Analysis of the Person-Response Function

Continuous estimates of PRFs were obtained using kernel smoothing. Graphical displays of increasing or locally increasing PRFs indicate which subsets of item scores disagree with the expected item scores. The graphical inspection of PRFs based on kernel smoothing was modeled after the estimation of IRFs using kernel smoothing (Douglas & Cohen, 2001; Habing, 2001; Molenaar & Sijtsma, 2000; Ramsay, 1991, 2000; Sijtsma & Molenaar, 2002).

Kernel Smoothing Estimation of the PRF

Estimating PRFs by means of kernel smoothing is based on the item ordering from easy to difficult; that is, by increasing $(1 - \pi_j)$. The estimate

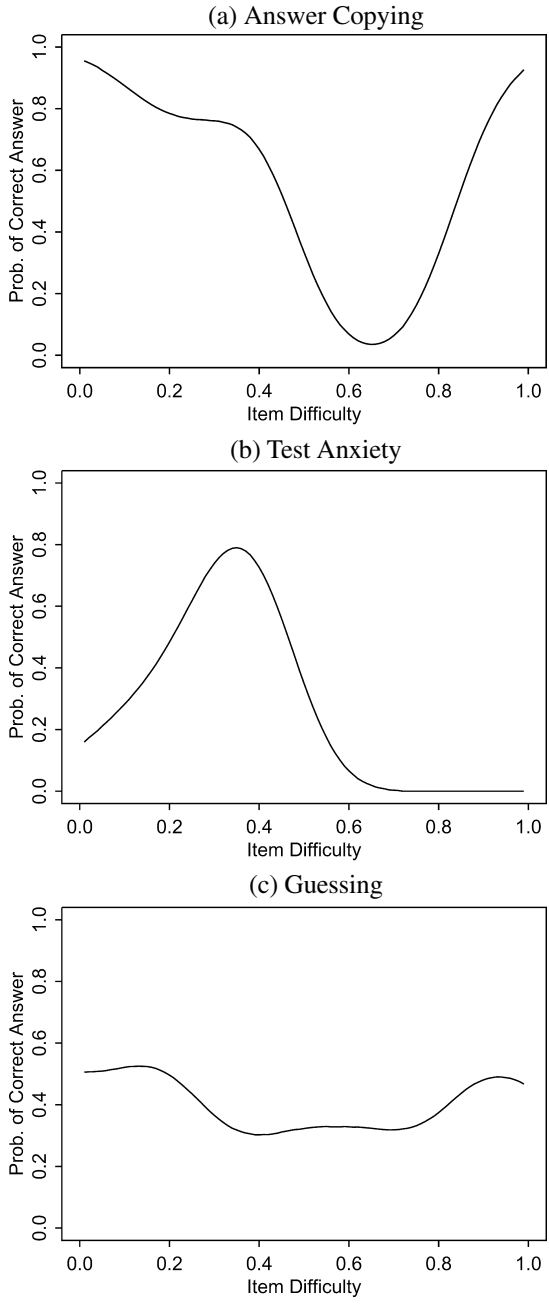


Figure 2
Hypothetical PRFs for (a) Answer Copying, (b) Test Anxiety, and (c) Guessing

of the PRF at focal point $(1 - \pi_0)$ is obtained by taking the weighted average of the scores of items with difficulty close to $(1 - \pi_0)$, and weights defined by the kernel function. The estimated function value $\hat{P}_v(1 - \pi_0)$ is given by

$$\hat{P}_v(1-\pi_0) = \sum_{j=1}^{J} \omega(\pi_j) x_{vj},$$

with weights

$$\omega(\pi_j) = \frac{K\left(\frac{\pi_0 - \pi_j}{h}\right)}{\sum_{j=1}^{J} K\left(\frac{\pi_0 - \pi_j}{h}\right)}.$$

Constant h is the bandwidth value for which a suitable value (to be explained shortly) is chosen by the researcher. We used the Gaussian kernel function,

$$K\left(\frac{\pi_0 - \pi_j}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left[-\left(\frac{\pi_0 - \pi_j}{h}\right)^2\right].$$

The maximum of $\omega(\pi_j)$ is found when $\pi_0 = \pi_j$, and $\omega(\pi_j)$ goes to zero as $|\pi_0 - \pi_j|$ increases. In practice, $\hat{P}_v(1 - \pi)$ is computed at a large number of focal points, to be specified by the user. Experience has shown that using 100 focal points yields accurate estimates of the PRF (e.g., Ramsay, 1991, used 51 equally spaced focal points to estimate IRFs).

The user-specified bandwidth h controls the trade-off between bias and sampling variation (e.g., Simonoff, 1996, pp. 22-23). Small h values yield estimated PRFs with large variance and small bias, and large h values yield estimated PRFs with small variance and large bias. For small h, the estimated PRFs may show several local deviations due to sampling variation, possibly resulting in many item-score vectors incorrectly identified as misfitting. However, for large h, important deviations of the observed PRF from the expected PRF may be overlooked and, as a result, only few item-score vectors may be classified as misfitting. Appropriate values for h were determined by trying a variety of h values and inspecting the estimated PRFs. A real data example to be discussed shortly illustrates the choice of h (also, see e.g., Douglas & Cohen, 2001; Habing, 2001). The h values that ignored small fluctuations in the PRF but traced large deviations were considered to be suitable choices. This heuristic approach also adapts the

choice of h to the number of items and their spread, and must be repeated for each data set. End-point bias, caused by a small number of data points available to estimate the tails of the PRFs (Ramsay, 1991; Simonoff, 1996, p. 49), was not believed to be a threat here, because most psychological and educational tests have π_js that are evenly distributed over the interval [0,1]. Thus, there are no extreme values that disproportionately influence the shape of the PRF.

PRF Variability Bands

Variability bands were constructed using a jackknife procedure (e.g., Efron & Tibshirani, 1993, pp. 141-150; see Bowman & Azzalini, 1997, p. 75; and Ramsay, 1991, for alternative methods). These bands express the uncertainty associated with the estimated PRF. To obtain the jackknife variability bands, we estimated the PRF J times, in each estimation round leaving out another item score X_j from the sample of J item scores of respondent υ. Let $\mathbf{X}_{(-j)}$ be the item-score vector omitting item score X_j. This is the j^{th} jackknife sample. Let the j^{th} jackknife estimate of the PRF at focal point $(1 - \pi_0)$ be denoted by $\hat{P}_\upsilon [1 - \pi_0; \mathbf{X}_{(-j)}]$. The jackknife estimate of the standard error (SE) at focal point π_0 is defined by

$$\widehat{\mathrm{SE}}_{\upsilon jack} = \left(\frac{J-1}{J} \sum_{j=1}^{J} \left\{ \hat{P}_\upsilon \left[1-\pi; \mathbf{X}_{(-j)} \right] - \hat{P}_{\upsilon jack} \left(1-\pi_0 \right) \right\}^2 \right)^{1/2}, \tag{2}$$

with

$$\hat{P}_{\upsilon jack} \left(1-\pi_0 \right) = J^{-1} \sum_{j=1}^{J} \hat{P}_\upsilon \left[1-\pi_0; \mathbf{X}_{(-j)} \right]. \tag{3}$$

The jackknife $\widehat{\mathrm{SE}}$s are calculated on the basis of the J estimates of $P_\upsilon(1 - \pi_0)$, and this is repeated for a large number of focal points $(1 - \pi_0)$. At each focal point $(1 - \pi_0)$, the $(1 - \alpha)$ variability band is given by $\hat{P}_\upsilon (1 - \pi_0) \pm z_\alpha \times \widehat{\mathrm{SE}}_{\upsilon jack}$.

Simulated Data Example of Kernel Smoothing of the PRF

A simulated data example illustrates the accuracy of PRF estimation using kernel smoothing. As an example, let a PRF be defined by

$$P_\upsilon(\delta) = \{1 + \exp[\alpha_\upsilon(\delta - \theta_\upsilon)]\}^{-1}, \tag{4}$$

with $\delta \in [0, 1]$, and fixed θ_v and α_v. For convenient choices $\theta_v = 0.5$ and $\alpha_v = 6$, and test length $J = 45$, response probabilities $P_v(\delta_j)$ ($j = 1, ..., 45$) were obtained from Equation 4 for 45 equidistant points on the δ scale. Next, 100 binary item-score vectors $\mathbf{X} = \mathbf{x}$ were simulated. Each item-score vector was generated by drawing for each item a score from a Bernoulli distribution with parameter $P_v(\delta_j)$. Then, for each of the 100 simulated vectors $\mathbf{X} = \mathbf{x}$, the PRF was estimated for $h = 0.05$, 0.09, and 0.13, respectively. This resulted in 100 replicated PRF estimates for each h. Let $\hat{P}_v(\delta_t)$ be the PRF estimate at focal point δ_t ($t = 1, ..., T$). The accuracy of a PRF estimate was evaluated using the mean of the sum of squared errors (*SSE*) between the estimated PRF, $\hat{P}_v(\delta_t)$, and the theoretical PRF, $P_v(\delta_t)$, across the T focal points; that is,

$$SSE = \sum_{t=1}^{T} \left[\hat{P}_v(\delta_t) - P_v(\delta_t) \right]^2 .$$

Let SSE_M be the mean of the *SSEs* resulting from the 100 replications. The bias of a PRF estimate was evaluated as follows. First, the squared difference between the mean PRF estimate across the 100 bootstrap samples at focal point (δ_t), denoted by $\hat{P}_v(\delta_t)_M$, and the theoretical PRF at δ_t (Equation 4) was determined for $t = 1, ..., T$. Then, these squared differences were added across the T focal points; that is,

$$BIAS = \sum_{t=1}^{T} \left[\hat{P}_v(\delta_t)_M - P_v(\delta_t) \right]^2 .$$

For $h = 0.05$, 0.09, and 0.13, Figures 3a through 3c, respectively, each shows the true PRF (solid curve; obtained from Equation 4), the mean of the PRF estimates (dashed curve), and five representative PRFs (dotted curves; their $SSE \approx SSE_M$). These representative PRFs give an indication of the mean accuracy of the estimated PRFs. For $h = 0.05$ ($SSE_M = 1.969$ and $BIAS = 0.019$), the PRF estimates are too inaccurate, leading to many Type I errors. For $h = 0.13$ ($SSE_M = 0.985$ and $BIAS = 0.109$), most of the sampling variation was smoothed away and PRF estimates tended to become linear (except for the tails). In this example, $h = 0.09$ ($SSE_M = 1.077$ and $BIAS = 0.025$) appears to represent a good compromise because the SSE_M closely resembles that for $h = 0.13$, while *BIAS* resembles that of $h = 0.05$.

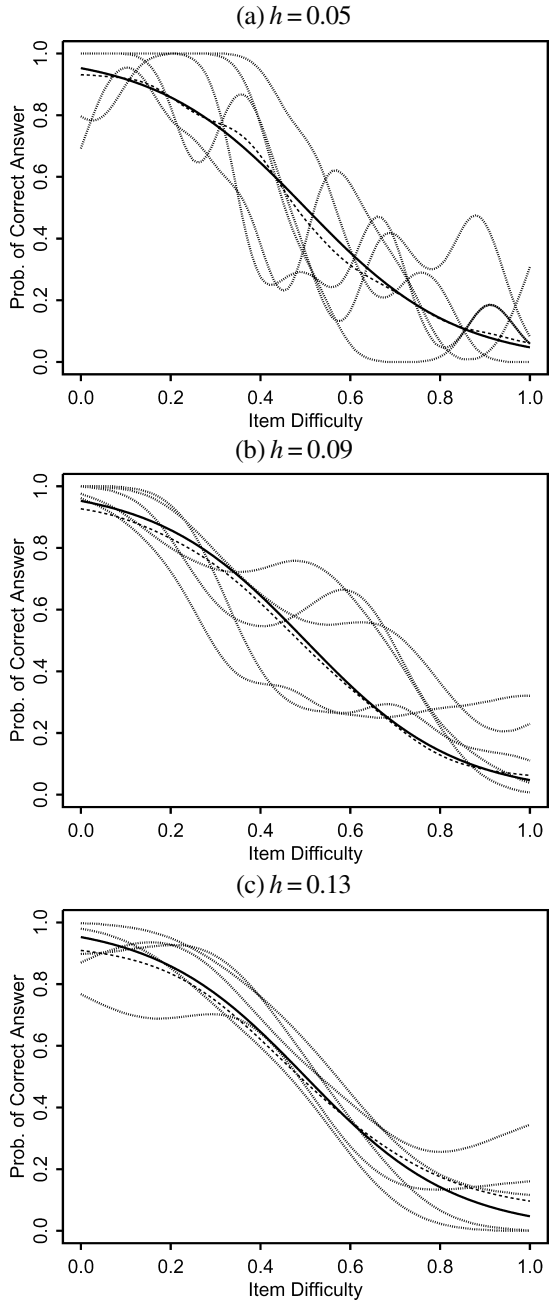


Figure 3
True PRF (Solid Line), Mean of Bootstrap PRF Estimates (Dashed Line), and Five Representative PRF Estimates (Dotted Lines) for Three Levels of h

Real Data Example of Kernel Smoothing of the PRF

A real data example was taken from the subscale Hidden Figures ($J = 45$), which is part of the Revised Amsterdam Child Intelligence Test (RAKIT; Bleichrodt, Drenth, Zaal, & Resing, 1984). Figures 4a through 4c show estimates of the same PRF with corresponding 90% variability bands, for bandwidth values $h = 0.05$, 0.09 and 0.13, respectively. For $h = 0.05$, the PRF estimate shows a small local increase at the items that were relatively easy in this test [$.35 \leq (1 - \pi) \leq .45$]. At these items, the variability bands indicate large sampling error, suggesting lack of evidence about misfit. A larger increase was found for $.60 \leq (1 - \pi) \leq .80$. This increase is larger than the width of the variability bands, implying convincing evidence of misfit. For $h = 0.09$ and $h = 0.13$, the first local increase was smoothed away, whereas the second local increase was visible. However, for $h = 0.13$ the data window used for smoothing produced estimates showing little variation, thus estimating an almost horizontal PRF.

These and other plots not shown here, suggest that h values between 0.07 and 0.11 are reasonable choices. This agrees with the simulation result. Finally, notice that the right tail of a PRF is estimated with great inaccuracy. This is due to greater variation in correct and incorrect answers to the more difficult items than to the easier items.

Person-Fit Assessment using Logistic Regression

Cheating on the most difficult items, for example, may result in a U-shaped PRF. A logistic regression model (e.g., Agresti, 1990; Fox, 1997; McCullagh, 1980) may fit this U-shape to the observed data and test the model against interesting alternatives, such as a monotone decreasing PRF. This may result in an increased power to detect cheating and other kinds of misfit. The estimated logistic regression parameters may suggest causes of misfit. Logistic regression may be a useful statistical tool additional to the visual inspection of a large number of PRF graphs.

Logistic Regression Models

Let item-score variable X_j be the response variable and item rank number j the discrete ordinal explanatory variable for the item difficulty. Let the log-odds of the PRF $P_v(1 - \pi_j)$ be a linear function of j, similar to the linear-by-linear model (Agresti, 1990, p. 265; McCullagh, 1980),

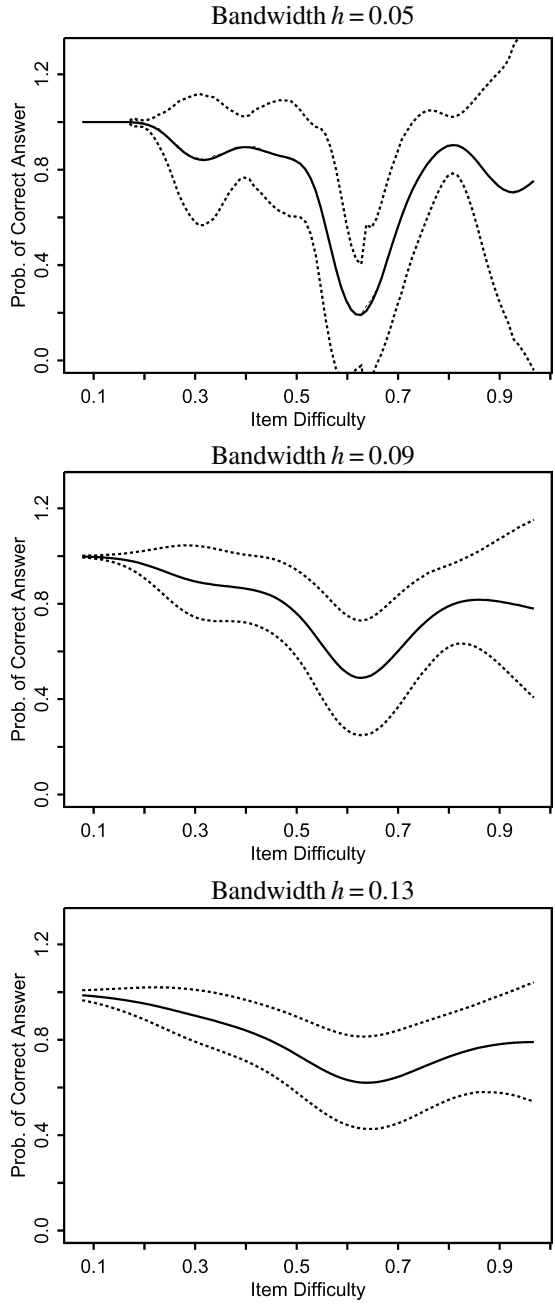


Figure 4
Estimated (Quasi)-Continuous Person-Response Functions for Different h Values

$$\log\left[\frac{P_v\left(1-\pi_j\right)}{1-P_v\left(1-\pi_j\right)}\right]=\alpha+\beta j. \quad (5)$$

IIO implies that $P_v(1-\pi_j)$ is nonincreasing as rank number j increases. Thus, in Equation 5 slope β must be zero or negative, whereas intercept α is unrestricted. Because in NIRT the item difficulty $(1-\pi)$ is treated as an ordinal variable, the magnitude of β is of little value. However, a positive β indicates that the mean trend of the PRF is positive, and this indicates misfit. Thus, we only use information about the sign of β. Notice that in a parametric IRT context, the magnitude of β may be interpreted as an index of an individual's measurement precision (Reise, 2000; Trabin & Weiss, 1983) and may be used in person-fit analysis (Strandmark & Linn, 1987).

First, the global trend of the PRF may be investigated by testing the null hypothesis that $\beta = 0$ (borderline case of fit) against the alternative that $\beta > 0$ (misfit). This is done using a likelihood ratio test, which compares the fit of the full model (Equation 5) and the restricted (null) model for $\beta = 0$; that is,

$$\log\left[\frac{P_v\left(1-\pi_j\right)}{1-P_v\left(1-\pi_j\right)}\right]=\alpha. \quad (6)$$

Let L_0 be the maximum likelihood for the null model (Equation 6) and L_1 the maximum likelihood for the full model (Equation 5). Then, the likelihood ratio test statistic for the null hypothesis $\beta = 0$ is $G^2_\beta = -2(\ln L_0 - \ln L_1)$, which follows a χ^2 distribution with $df = 1$.

Second, the curvature of U-shaped or bell-shaped PRFs may be captured by a quadratic logistic regression model,

$$\log\left[\frac{P_v\left(1-\pi_j\right)}{1-P_v\left(1-\pi_j\right)}\right]=\alpha+\beta j+\gamma j^2. \quad (7)$$

For $\gamma > 0$, Equation 7 describes a U-shaped PRF, which indicates misfit associated with spuriously high X_+ scores. For $\gamma < 0$, Equation 7 describes a bell-shaped PRF, indicating misfit associated with spuriously low X_+ scores.

To detect U-shaped PRFs, null hypothesis $\gamma = 0$ is tested against the alternative that $\gamma > 0$. This is done by testing the full quadratic model (Equation 7) against the linear null model (Equation 5). Let L_1 and L_0 denote the maximum likelihood of the full model and the null model, respectively.

Under the null hypothesis $\gamma = 0$, the test statistic $G^2_\gamma = -2(\ln L_0 - \ln L_1)$ has an asymptotic χ^2 distribution with $df = 1$.

Application of Logistic Regression Models to Person-Fit Analysis

The logistic regression model is used to test global misfit, and misfit associated with spuriously low or high X_+ scores.

1. *Detection of global person misfit.* Parameter β indicates the linear main trend of the PRF (Equation 5). A positive β gives evidence of PRF misfit without locating this misfit at the item difficulty scale. Positive $\hat{\beta}$s may be selected and tested for significance, using the likelihood ratio test statistic G^2_β.

2. *Detection of spuriously high X+ scores* (sp – HS). A meaningful test of a U-shaped PRF has to satisfy two conditions: (a) a test of linear main trend β reveals that the slope is nonnegative and (b) the test that $\gamma = 0$ against $\gamma > 0$ yields a log-likelihood ratio $G^2_\gamma \geq 2.71$ [χ^2 test with $df = 1$ and $\alpha = .05$; from now on, test is denoted G^2_γ (sp – HS)]. Condition 1 prevents that fitting PRFs showing a significant positive quadratic effect (ie., convex upwards but monotone decreasing) are incorrectly flagged as misfitting.

3. *Detection of spuriously low scores* (sp – LS). A meaningful test of a bell-shaped PRF has to satisfy two conditions: (a) a test of linear main trend β reveals that the slope is nonnegative; and (b) the test that $\gamma = 0$ against $\gamma < 0$ yields a log-likelihood ratio $G^2_\gamma \geq 2.71$ [from now on, test is denoted G^2_γ (sp – LS)]. Condition 1 prevents that fitting PRFs having a negative quadratic effect (i.e., convex downwards and monotone decreasing) are incorrectly flagged as misfitting.

Examples of Person-Fit Analysis using Logistic Regression

This section illustrates person-fit analysis using logistic regression for six hypothetical item score vectors ($J = 20$; Table 1). Figure 5 shows the estimated PRFs and their variability bands. For Cases 1 and 2, the PRFs do not show local increases, but they differ in shape; that is, the PRF for Case 1 follows a logistic curve and the PRF for Case 2 a near straight line. The PRF for Case 3 has a U-shape. The PRF for Case 4 shows an increase but is not U-shaped or bell-shaped. The PRFs for Cases 5 and 6 have a bell-shape.

Table 1 shows for each item-score vector the $\hat{\beta}$s from the linear model (Equation 5), the $\hat{\gamma}$s from the quadratic model (Equation 7), and the results of the likelihood-ratio test for the parameters. Based on the test results, it was concluded that $\beta < 0$ for Cases 1, 2, and 5; and that $\beta = 0$ for Cases 3, 4, and 6. Thus, none of the tests led to the conclusion of a positive linear

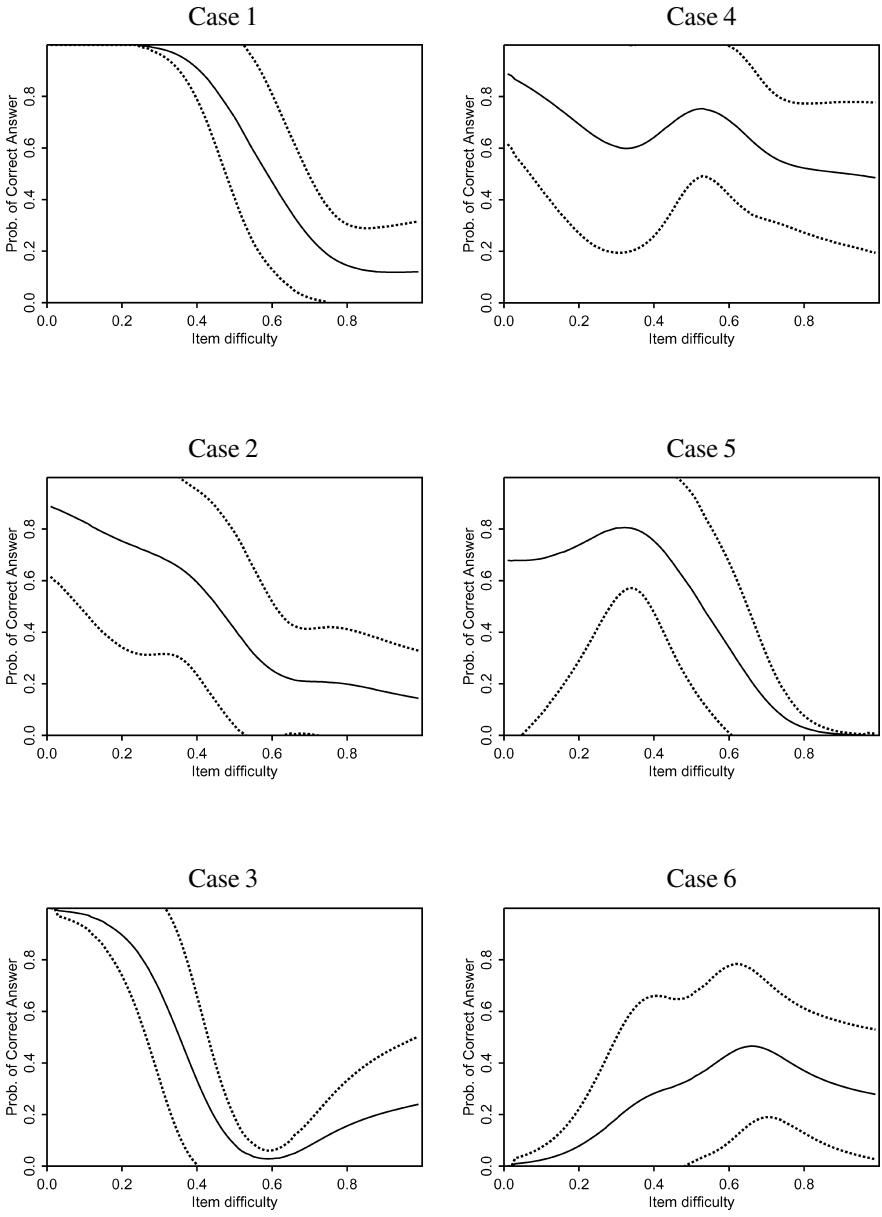


Figure 5
Person-Response Functions for Six Hypothetical Item-Score Vectors

trend. Also, it was concluded that $\gamma > 0$ (convex upwards) for Cases 1 and 3; and that $\gamma < 0$ (convex downwards) for Cases 4 and 6.

For Case 1, the combination of a negative β (Table 1) and a monotone decreasing PRF (Figure 5), and a positive γ (Table 1), which suggests convexity upwards, provides evidence of person fit. For Case 2, there is evidence of a linear trend downwards, suggesting person fit. For Case 3, a zero β in combination with a positive γ, suggesting convexity upwards, provides evidence of person misfit. For Case 4, there is no linear trend but there is evidence of convexity downwards. The graph does not show a clear-cut bell-shape. Combining these results, for Case 4 the situation is not clear. For Case 5, the negative β indicates a negative mean trend in the PRF. In Figure 5, the variability bands show that the small increase of the PRF at the first items should not be taken seriously. The negative γ, which suggests convexity downwards, does not alter this conclusion. For Case 6, it was concluded that $\beta = 0$, meaning that no significant linear mean trend was found. The result that $\gamma < 0$ means that the PRF is convex downwards, and thus flags Case 6 as misfitting.

Comparison with Van der Flier's U3 Statistic

The methods presented in this article were compared with the NIRT person-fit statistic *U*3 (Van der Flier, 1980, 1982). Statistic *U*3 expresses the degree to which an individual item-score vector deviates from the item-score vectors of the majority of respondents in the sample. For the ordered item-score vector **X**, statistic *U*3 equals

Table 1
Six Hypothetical Item-Score Vectors and Corresponding Estimated Parameters of the Logistic Regression Model and Likelihood Ratio Statistics

Case	Item-Score Vector				*ZU*3	$\hat{\beta}$	G^2_{β}	$\hat{\gamma}$	G^2_{γ}
1	11111	11001	00000	00100	–1.05	–.3544	10.82	.0659	3.75
2	11011	01000	10010	00000	–.25	–.2511	6.47	–.0001	.00
3	11110	00000	00000	00110	1.28	–.1234	1.90	.0917	12.06
4	10011	10110	11111	10000	1.06	–.0774	.92	–.0388	4.79
5	10111	10100	00000	00000	–.45	–.5263	11.30	–.1622	1.77
6	00001	00011	01100	00100	2.66	.0288	.11	–.0384	3.41

$$U3(\mathbf{X}) = \frac{\Sigma_{j=1}^{X_+} \text{logit}(\pi_j) - \Sigma_{j=1}^{J} X_j \, \text{logit}(\pi_j)}{\Sigma_{j=1}^{X_+} \text{logit}(\pi_j) - \Sigma_{j=X_+ +1}^{J} \text{logit}(\pi_j)}.$$

Statistic $U3$ equals 0 if the correct answers are in the first X_+ positions (i.e., the X_+ easiest items) of $\mathbf{X}$ and $U3$ equals 1 if the correct answers are in the last X_+ positions (i.e., the X_+ most difficult items). Van der Flier (1980, 1982) proposed a standardized version of $U3$, denoted $ZU3$, which has an asymptotic standard normal distribution for low to medium item discrimination power (corresponding to slope parameter values of logistic IRFs ranging from .5 to 1.5; to be defined shortly). For tests with high item discrimination, the significance test for $ZU3$ is not suitable (Emons, Meijer, & Sijtsma, 2002). However, in this case $U3$ can be used descriptively to order item-score vectors according to their likelihood. In practice, the researcher may proceed by selecting, for example, the five percent of the item-score vectors that have the highest $U3$ values. In this study, we explored whether the test on the β parameter from the logistic regression model provides a useful alternative to the $U3$ statistic. An advantage of $\hat{\beta}$ is its easy use: a significantly positive $\hat{\beta}$ indicates misfit.

Simulation Study

A simulation study was done to investigate the usefulness of logistic regression models for person-fit analysis. False alarm rates and detection rates were investigated for the test on the β regression parameter for linear trend and the $U3$ test, and for the test on the γ regression parameter for convexity upwards and convexity downwards of the PRFs.

Method

Data Simulation and Model Estimation

Item-score vectors were simulated using the flexible four-parameter logistic model (4-PLM; Hambleton & Swaminathan, 1985, p. 48),

$$P_j(\theta) = \zeta_j + (\lambda_j - \zeta_j) \frac{\exp\left[D\alpha_j(\theta - \delta_j)\right]}{1 + \exp\left[D\alpha_j(\theta - \delta_j)\right]}, \quad (8)$$

where ζ_j is the lower asymptote for $\theta \rightarrow -\infty$, λ_j the upper asymptote for $\theta \rightarrow \infty$, α_j is the discrimination parameter, δ_j the location parameter, and D

= 1.7 is the scaling factor. The parameters for the 4-PLM were chosen such that the IRFs did not intersect (e.g., Sijtsma & Meijer, 2001). Logistic regression models were fitted to each item-score vector using the computer program ℓEM (Vermunt, 1997), under the assumption that **X** follows a product binomial distribution.

Independent Variables

The following variables were manipulated:

1. *Test Length*. Data were simulated for two levels of test length: $J = 20$ and $J = 40$.

2. *Item Discrimination Power*. For both levels of test length, two set-ups for the α_j parameters were used, resulting in one set of IRFs with low discrimination ($\alpha = 1$) and one set of IRFs with high discrimination ($\alpha = 2$).

3. *Aberrant Response Behavior*. Two types of aberrant response behavior were simulated. The first one was *Answer Copying*. Item scores were simulated under the 4-PLM, but for difficult items $P_j(\theta)$ was a priori fixed to 1.00. This resulted in spuriously high X_+ scores. The second one was *Test Anxiety*. Item scores were simulated under the 4-PLM, but for easy items $P_j(\theta)$ was a priori fixed to .25. This resulted in spuriously low X_+ scores.

4. *Number of Items Exhibiting Misfit*. For both test lengths, two levels of the number of misfitting items, denoted by J_m, were simulated. For the 20-item test, $J_m = 5, 8$; and for the 40-item test, $J_m = 5, 10$.

5. θ-*level*. For each level of test length, item discrimination and type of misfit, item score vectors were simulated at three θ-levels: (a) θ randomly drawn from $N(-1, 0.5)$ [*Low*]; (b) θ randomly drawn from $N(0, 0.5)$ [*Medium*]; and (c) θ randomly drawn from $N(1, 0.5)$ [*High*]. At each θ-level, 1000 item-score vectors were simulated.

The result is a cross-factorial design with $2 \times 2 \times 2 \times 2 \times 3 = 48$ cells.

Dependent Variables

The dependent variables were the false alarm rates and the detection rates. They were evaluated separately for the linear trend test on regression coefficient β, G^2_β, the tests on γ for convexity upwards [spuriously high X_+ scores; test G^2_γ(sp – HS)] and convexity downwards [spuriously low X_+ scores; test G^2_γ(sp – LS)], and Van der Flier's *ZU*3 statistic. We also investigated the joint false alarm rates and the joint detection rates when at

least one of the three tests on β and γ was significant. In addition, because its sign is informative of the slope of the linear trend of the PRF, $\hat{\beta}$ was used as a descriptive statistic. Note that detection rates for $\hat{\beta}$ are at least as large as those for the test on β, G_{β}^2. Item-score vectors with $X_+ \leq 2$ or $X_+ \geq J-2$ were excluded from the analysis, because they contained too little information for useful person-fit analysis. All significance tests were done at the five percent significance level.

Results of the Simulation Study

False Alarm Rates

Test Length. For $J = 20$ (Table 2, upper half), the false alarm rates for statistic $\hat{\beta}$ and the logistic regression PRF tests [i.e, G_{β}^2, G_{γ}^2 (sp – HS), and G_{γ}^2 (sp – LS)] ranged from .000 to .041 (Table 2, Columns 3 through 6). The joint false alarm rates ranged from .017 and .041 (Table 2 , Column 7). The false alarm rates for *ZU*3 (Table 2, Column 8) were comparable to those for the joint tests or they were lower. For the logistic regression PRF tests, the false alarm rates were lower for $J = 40$ (Table 2, lower half) than for $J = 20$. Thus, these tests were conservative, and for larger J they were more conservative. The explanation probably is that under the null-model all item-score vectors represented decreasing PRFs, and only because of sampling error may a PRF become flatter (but still decreasing), or increasing or locally increasing. Only the latter cases are candidates for significance testing. Thus, the Type I error is expected to be smaller than the nominal level, and the effect is stronger for larger samples (i.e., larger J). Compared with $J = 20$, for $J = 40$ the false alarm rates for *ZU*3 were lower for medium θ, but higher for low and high θ, even exceeding the nominal significance level.

Item Discrimination. For the logistic regression PRF tests, the effects of increasing item discrimination were comparable with those of increasing test length. That is, higher item discrimination resulted in lower false alarm rates.

Test Length × *Item Discrimination*. For none of the person-fit statistics interaction effects were found.

Table 2
False Alarm Rates for the Logistic Regression Person-Fit Tests and Van der Flier's ZU3 Statistic

		Person-Fit Tests					
θ	*n*	$\hat{\beta}$	G^2_{β}	G^2_{γ} (sp – HS)	G^2_{γ} (sp – LS)	Joint	*ZU*3
		20-Item Test, Low Item Discrimination					
Low	986	.007	.000	.005	.018	.023	.024
Medium	1000	.001	.000	.005	.012	.017	.002
High	947	.001	.000	.006	.018	.024	.019
Overall	2933	.003	.000	.006	.016	.022	.015
		20-Item Test, High Item Discrimination					
Low	977	.004	.001	.010	.006	.017	.016
Medium	997	.000	.000	.001	.003	.004	.000
High	928	.004	.000	.000	.041	.041	.010
Overall	2902	.003	.000	.004	.016	.020	.009
		40-Item Test, Low Item Discrimination					
Low	999	.002	.000	.002	.003	.005	.132
Medium	1000	.000	.000	.000	.000	.000	.014
High	998	.000	.000	.000	.008	.008	.129
Overall	2997	.000	.000	.001	.004	.004	.087
		40-Item Test, High Item Discrimination					
Low	1000	.001	.000	.001	.003	.000	.200
Medium	1000	.000	.000	.000	.000	.000	.013
High	995	.001	.000	.001	.009	.010	.174
Overall	2995	.001	.000	.001	.004	.005	.123

Note. $\hat{\beta}$ is the descriptive global person-fit statistic; G^2_{β} denotes the test whether $\hat{\beta}$ is significantly greater than 0; G^2_{γ} (sp – HS) denotes the test for detection of spuriously high X_+ scores; G^2_{γ} (sp – LS) denotes the test for detection of spuriously low X_+ scores; and *Joint* denotes that at least one of G^2_{β}, G^2_{γ} (sp – HS), and G^2_{γ} (sp – LS) was significant.

Detection Rates

Results for Answer Copying

Comparison of Person-fit Methods. The detection rates for $\hat{\beta}$ were considerably higher than those for G^2_β (Tables 3 and 4, Columns 3 and 4), except for $J = 40$ and $J_m = 5$. For example, for $J = 20$ and $J_m = 5$, the detection rates for $\hat{\beta}$ ranged from .18 to .68, whereas the detection rates using G^2_β ranged from .00 to .08. For $J = 20$ and $J_m = 8$, $\hat{\beta}$ detected at least 89% of the aberrant item-score vectors, but only 5% through 31% of the item-score vectors yielded a significant G^2_β. For $J = 40$, and $J_m = 5$, $\hat{\beta}$ yielded detection rates ranging from .00 to .19, whereas G^2_β yielded detection rates of at most .01. It may be concluded that G^2_β has too little power to detect PRF misfit.

Except for $J = 40$ and $J_m = 5$, G^2_γ(sp – HS) (Tables 3 and 4, Column 5) yielded detection rates ranging from .55 to .99, whereas the detection rates for G^2_γ(sp – LS) were zero for all levels of test length, item discrimination, number of misfitting items, and θ (Tables 3 and 4, Column 6). Comparison of G^2_γ(sp – HS) and $\hat{\beta}$ revealed that G^2_γ(sp – HS) yielded the highest detection rates, except for $J = 20$ and $J_m = 8$. Finally, the highest detection rates were found for *ZU*3, except for the combination of $J = 20$, $J_m = 5$, and high item discrimination, where G^2_γ(sp – HS) yielded the highest detection rates on average. It may be noted that for $J = 40$ and $J_m = 10$, high detection rates for *ZU*3 for low and high θs go together with false alarm rates larger than the 5% nominal significance level (see Table 2, Column 8).

Item Discrimination Power. The detection rates for $\hat{\beta}$ and G^2_β were lower for high item discrimination than for low item discrimination (Tables 3 and 4; Columns 3 and 4); the smallest differences were found for $J = 20$ with $J_m = 8$, and the largest differences were found for $J = 40$ with $J_m = 10$. Compared with detection rates for low item discrimination, for high item discrimination the detection rates for G^2_γ(sp – HS) were higher for $J = 20$, but equal or lower for $J = 40$.

Test Length. Comparison of the detection rates for $\hat{\beta}$ and G^2_β for $J = 20$ and $J_m = 5$ with those for $J = 40$ and $J_m = 10$ (i.e, 25 percent misfit in both cases) showed positive and negative differences; absolute differences ranged from .00 to .08. For G^2_γ(sp – HS) and low item discrimination, the detection rates for $J = 40$ and $J_m = 10$ were higher than for $J = 20$ and $J_m = 5$; differences in detection rates ranged from .08 to .14.

Table 3
Detection Rates for the Logistic Regression Person-Fit Tests and Van der Flier's *ZU*3 statistic, for Answer Copying and $J = 20$

		Person-Fit Tests					
θ	n	$\hat{\beta}$	G^2_β	G^2_γ (sp – HS)	G^2_γ (sp – LS)	Joint	*ZU*3
		Five Items Misfit, Low Discrimination					
Low	990	.68	.08	.83	.00	.87	.96
Medium	885	.37	.02	.77	.00	.78	.87
High	475	.33	.00	.62	.00	.63	.86
Overall	2350	.50	.04	.77	.00	.79	.91
		Five Items Misfit, High Discrimination					
Low	990	.56	.07	.93	.00	.95	.93
Medium	837	.18	.00	.86	.00	.86	.72
High	356	.24	.01	.68	.00	.69	.72
Overall	2183	.37	.04	.86	.00	.87	.81
		Eight Items Misfit, Low Discrimination					
Low	924	.99	.31	.81	.00	.92	1.00
Medium	572	.94	.12	.73	.00	.81	1.00
High	169	.94	.09	.55	.00	.62	1.00
Overall	1665	.97	.24	.76	.00	.85	1.00
		Eight Items Misfit, High Discrimination					
Low	906	.96	.23	.92	.00	.95	1.00
Medium	442	.89	.05	.80	.00	.83	1.00
High	75	.93	.08	.57	.00	.65	1.00
Overall	1423	.94	.17	.86	.00	.90	1.00

Note. $\hat{\beta}$ is the descriptive global person-fit statistic; G^2_β denotes the test whether $\hat{\beta}$ is significantly greater than 0; G^2_γ (sp – HS) denotes the test for detection of spuriously high X_+ scores; G^2_γ (sp – LS) denotes the test for detection of spuriously low X_+ scores; and *Joint* denotes that at least one of G^2_β, G^2_γ (sp – HS), and G^2_γ (sp – LS) was significant.

Table 4
Detection Rates for the Logistic Regression Person-Fit Tests and Van der Flier's *ZU*3 statistic, for Answer Copying and $J = 40$

		Person-Fit Tests					
θ	n	$\hat{\beta}$	G^2_{β}	G^2_{γ} (sp – HS)	G^2_{γ} (sp – LS)	Joint	*ZU*3
			Five Items Misfit, Low Discrimination				
Low	1000	.19	.01	.73	.00	.73	.98
Medium	1000	.01	.00	.32	.00	.33	.92
High	977	.03	.00	.21	.00	.21	.97
Overall	2977	.08	.00	.42	.00	.42	.96
			Five Items Misfit, High Discrimination				
Low	1000	.13	.01	.61	.00	.61	1.00
Medium	1000	.00	.00	.12	.00	.12	.99
High	946	.04	.00	.06	.00	.06	1.00
Overall	2946	.06	.00	.26	.00	.26	.99
			Ten Items Misfit, Low Discrimination				
Low	1000	.76	.15	.97	.00	.98	1.00
Medium	993	.37	.01	.89	.00	.90	1.00
High	879	.32	.00	.70	.00	.71	1.00
Overall	2872	.49	.06	.86	.00	.87	1.00
			Ten Items Misfit, High Discrimination				
Low	1000	.59	.13	.99	.00	.99	1.00
Medium	978	.15	.00	.90	.00	.90	1.00
High	717	.22	.00	.67	.00	.67	1.00
Overall	2695	.33	.05	.87	.00	.87	1.00

Note. $\hat{\beta}$ is the descriptive global person-fit statistic; G^2_{β} denotes the test whether $\hat{\beta}$ is significantly greater than 0; G^2_{γ} (sp – HS) denotes the test for detection of spuriously high X_+ scores; G^2_{γ} (sp – LS) denotes the test for detection of spuriously low X_+ scores; and *Joint* denotes that at least one of G^2_{β}, G^2_{γ} (sp – HS), and G^2_{γ} (sp – LS) was significant.

Similar results were found for high item discrimination. Thus, test length had a small effect on the detection rates for the global person-fit test, and a somewhat larger effect on the detection rates for G_γ^2 (sp – HS).

Number of Items Exhibiting Misfit. For $J = 20$, the detection rates for $\hat{\beta}$ were higher for a larger number of misfitting items; differences in detection rates ranged from .31 to .61 for low item discrimination, and from .40 to .71 for high item discrimination. In addition, the detection rates for G_β^2 test were also higher for $J_m = 8$ than for $J_m = 5$, but these differences were smaller than those for $\hat{\beta}$. For G_γ^2 (sp – HS), the detection rates were a little lower for $J_m = 8$ than for $J_m = 5$. For $J = 40$, the results showed that if J_m is small, almost none of the tests on regression parameters yielded high detection rates; detection rates were smaller than .32 for medium and high θ levels.

For $J_m = 5$, G_γ^2 (sp – HS) yielded higher detection rates than $\hat{\beta}$. As the number of items showing misfit is larger, misfit manifests itself more as a linear trend picked up by the linear trend β parameter and less by the quadratic trend γ parameters. To illustrate this effect, we computed detection rates for answer copying, for $J = 40$, and 8, 12, 15, or 20 misfitting items. Figure 6 shows the detection rates for $J_m = 5$, 8, 10, 12, 15, and 20. The detection rates for $\hat{\beta}$ (solid curves) increased with increasing J_m and the detection rates for G_γ^2 (sp – HS) (dotted curves) first increased with J_m, and then decreased.

Detection Rates for Fixed 5% Type I Error Rates. The detection rates of the person-fit tests in Tables 3 and 4 were based on varying Type I error rates (Table 2). Thus, they may give a distorted picture of the absolute power of the statistics, making the comparison of detection rates less straightforward. For Answer Copying and $J = 20$, Table 5 shows the detection rates using fixed 5% Type I error rates. These detection rates were obtained using a critical value for each test statistic that corresponds to the 95th percentile of the sampling distribution underlying $J = 20$ (Table 2, upper panel).

For low and high item discrimination and all misfit levels, the detection rate for $\hat{\beta}$ increased to 1. However, the detection rate for G_β^2 ranged from .001 to .222 for $J_m = 5$, and from .070 to .591 for $J_m = 8$. Thus, for a fixed 5% Type I error rate statistic G_β^2 still had little power to detect misfit. For G_γ^2 (sp – HS) the detection rate was at least .95 for all misfit and item discrimination levels. The detection rate for G_γ(sp – LS) increased a little but was always smaller than .05. Thus, for a fixed 5% Type I error rate, the power of G_γ^2 (sp – HS) improved considerably, but G_γ^2 (sp – LS) remained insensitive to misfit. The joint detection rates

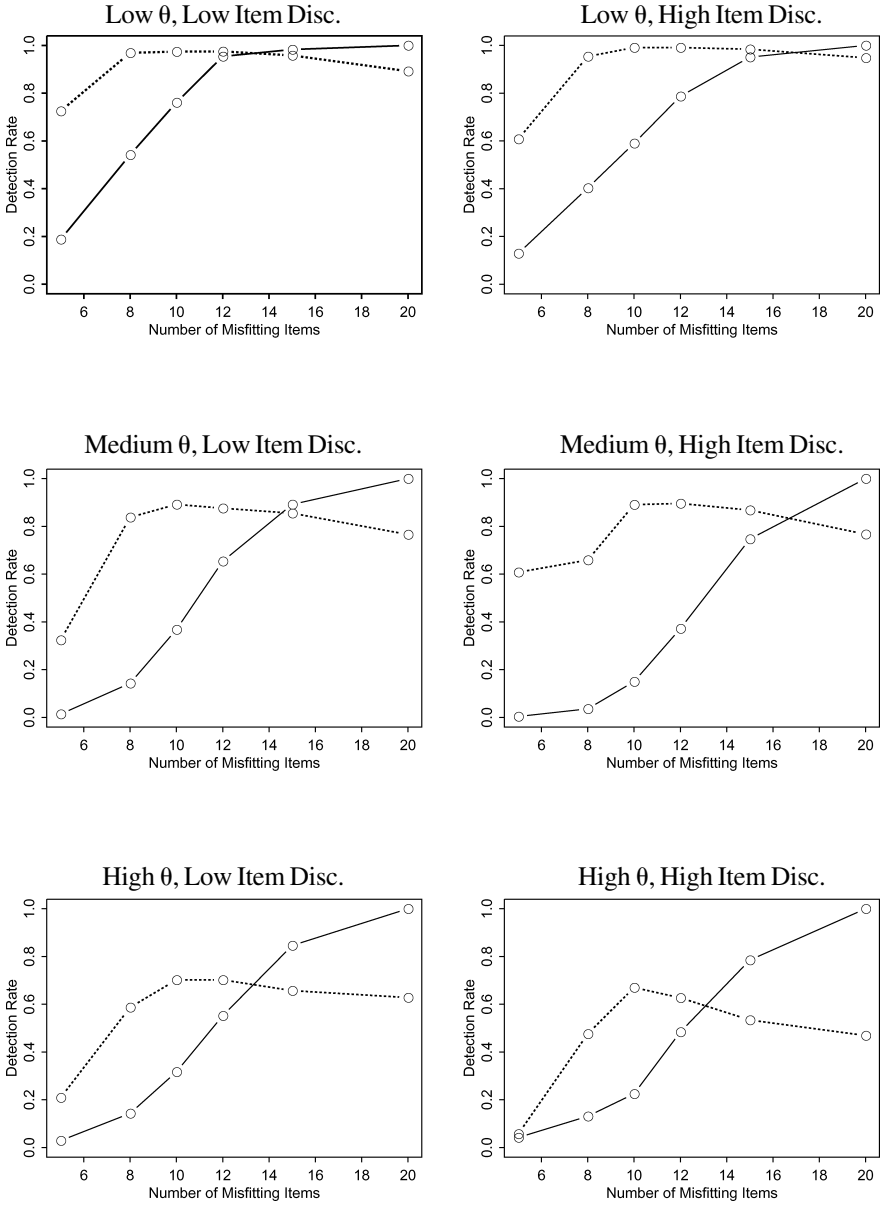


Figure 6

Detection Rates for $\hat{\beta}$ (Solid Curves) and G^2_γ (sp – HS) (Dotted Curves) for $J = 40$, and 5,8, 10, 12, 15, and 20 Misfitting Items, for Low and High Item Discrimination, and Three θ-Levels

Table 5
Detection Rates for the Logistic Regression Person-Fit Tests and Van der Flier's $U3$ Statistic, for Answer Copying, $J = 20$, and Fixed Type I Error Rate

		Person-Fit Tests					
θ	n	$\hat{\beta}$	G^2_β	G^2_γ (sp – HS)	G^2_γ (sp – LS)	Joint	$U3$
		Five Items Misfit, Low Discrimination					
Low	990	1.000	.222	.997	.003	1.000	.971
Medium	885	1.000	.049	.991	.009	1.000	.940
High	475	1.000	.050	.954	.037	.992	.895
Overall	2350	1.000	.120	.986	.012	.998	.936
		Five Items Misfit, High Discrimination					
Low	990	1.000	.093	.999	.001	1.000	1.000
Medium	837	1.000	.001	.995	.005	1.000	1.000
High	356	1.000	.014	.958	.039	.997	1.000
Overall	2183	1.000	.049	.991	.009	.999	1.000
		Eight Items Misfit, Low Discrimination					
Low	924	1.000	.591	.996	.003	.999	.964
Medium	572	1.000	.348	.991	.009	1.000	.901
High	169	1.000	.249	.982	.018	1.000	.841
Overall	1665	1.000	.472	.993	.006	.999	.909
		Eight Items Misfit, High Discrimination					
Low	906	1.000	.276	.999	.001	1.000	1.000
Medium	442	1.000	.070	.993	.007	1.000	1.000
High	75	1.000	.133	.973	.027	1.000	1.000
Overall	1423	1.000	.204	.996	.004	1.000	1.000

Note. $\hat{\beta}$ is the descriptive global person-fit statistic; G^2_β denotes the test whether $\hat{\beta}$ is significantly greater than 0; G^2_γ (sp – HS) denotes the test for detection of spuriously high X_+ scores; G^2_γ (sp – LS) denotes the test for detection of spuriously low X_+ scores; and *Joint* denotes that at least one of G^2_β, G^2_γ (sp – HS), and G^2_γ (sp – LS) was significant.

were close to 1. The detection rate for $U3$ was a little higher for low discrimination, and approximately the same for high discrimination. All results for $J = 40$ (not tabulated) were similar to those for $J = 20$.

Results for Test Anxiety

Compared with Answer Copying, the detection rates for Test Anxiety (Tables 6 and 7) were always smaller. Further, for Answer Copying the highest detection rates were found for low θ, and for Test Anxiety the highest detection rates were found for high θ. These results are consistent with other simulation studies (e.g., Meijer, Molenaar, & Sijtsma, 1994; Meijer & Sijtsma, 2001). A comparison of the effects of test length, item discrimination, and number of misfitting items on the detection rate for Answer Copying and that for Test Anxiety, in general led to similar conclusions, but the size of the effects was smaller for Test Anxiety than for Answer Copying.

Finally, detection rates were obtained for a fixed Type I error rate (not tabulated). The trends were comparable to those found for Answer Copying. In particular, the detection rate of $G^2_\gamma(\text{sp} - \text{LS})$ was greater than .84, whereas the detection rate for $G^2_\gamma(\text{sp} - \text{HS})$ was smaller than .16.

Conclusions and Discussion

Graphical PRF analysis using variability bands may be used to identify PRFs that exhibit deviations from monotone nonincreasingness. Logistic regression may then be used for modeling the PRF. The shape of the estimated PRF and the fitted logistic regression polynomial may help to evaluate the causes of misfit.[1] Here, we limited the discussion to quadratic U-shaped and bell-shaped logistic regression models. More complex polynomials may be used in principle to model other aberrant PRF shapes. However, preliminary calculations suggested that realistic test length may be too small to estimate and evaluate such complex models successfully.

In principle, our kernel smoothing methods and logistic regression methods can be used for investigating all causes of misfit that manifest themselves in deviations from the expected nonincreasingness of the PRF. Also, because test length J is limited for realistic and relevant tests, causes have to manifest themselves on several items to become sufficiently visible. This is a general problem in small-sample research as person-fit analysis typically is.

[1] The interested reader may contact us to obtain our software and instructions for its use. Modern freeware, such as ARC for graphical regression, could also be easily modified to produce person-response functions that can be explored via smoothing (e.g., Hart, 1997).

Table 6
Detection Rates for the Logistic Regression Person-Fit Tests and Van der Flier's *ZU*3 Statistic, for Test Anxiety and $J = 20$

		Person-Fit Tests					
θ	n	$\hat{\beta}$	G^2_{β}	G^2_{γ} (sp – HS)	G^2_{γ} (sp – LS)	Joint	*ZU*3
Five Items Misfit, Low Discrimination							
Low	917	.12	.01	.00	.34	.34	.26
Medium	990	.17	.00	.00	.55	.55	.28
High	1000	.42	.03	.00	.72	.74	.53
Overall	2907	.24	.01	.00	.54	.55	.36
Five Items Misfit, High Discrimination							
Low	844	.10	.00	.00	.29	.29	.23
Medium	986	.08	.00	.00	.58	.58	.15
High	1000	.33	.05	.00	.80	.82	.42
Overall	2830	.17	.02	.00	.57	.58	.27
Eight Items Misfit, Low Discrimination							
Low	835	.18	.00	.00	.26	.27	.36
Medium	953	.36	.02	.00	.41	.42	.53
High	997	.63	.11	.00	.52	.57	.75
Overall	2785	.40	.05	.00	.41	.43	.56
Eight Items Misfit, High Discrimination							
Low	771	.16	.00	.01	.25	.26	.23
Medium	930	.21	.00	.00	.47	.48	.35
High	997	.55	.10	.00	.66	.70	.66
Overall	2698	.32	.04	.00	.48	.50	.46

Note. $\hat{\beta}$ is the descriptive global person-fit statistic; G^2_{β} denotes the test whether $\hat{\beta}$ is significantly greater than 0; G^2_{γ} (sp – HS) denotes the test for detection of spuriously high X_+ scores; G^2_{γ} (sp – LS) denotes the test for detection of spuriously low X_+ scores; and *Joint* denotes that at least one of G^2_{β}, G^2_{γ} (sp – HS), and G^2_{γ} (sp – LS) was significant.

Table 7
Detection Rates for the Logistic Regression Person-Fit Tests and Van der Flier's *ZU*3 Statistic, for Test Anxiety and $J = 40$

		Person-Fit Tests					
θ	n	$\hat{\beta}$	G^2_{β}	G^2_{γ} (sp – HS)	G^2_{γ} (sp – LS)	Joint	*ZU*3
			Five Items Misfit, Low Discrimination				
Low	1000	.01	.00	.00	.12	.12	.50
Medium	1000	.00	.00	.00	.17	.17	.29
High	1000	.10	.00	.00	.48	.48	.58
Overall	3000	.03	.00	.00	.26	.26	.45
			Five Items Misfit, High Discrimination				
Low	996	.01	.00	.00	.06	.06	.24
Medium	1000	.00	.00	.00	.05	.05	.01
High	1000	.08	.01	.01	.39	.39	.09
Overall	2996	.03	.00	.00	.17	.17	.12
			Ten Items Misfit, Low Discrimination				
Low	997	.04	.00	.00	.33	.33	.80
Medium	1000	.08	.00	.00	.58	.58	.77
High	1000	.36	.00	.00	.85	.85	.92
Overall	2997	.16	.01	.00	.59	.59	.83
			Ten Items Misfit, High Discrimination				
Low	992	.04	.00	.00	.22	.22	.95
Medium	999	.03	.00	.00	.44	.44	.89
High	1000	.27	.05	.00	.79	.79	.95
Overall	2991	.11	.02	.00	.48	.48	.93

Note. $\hat{\beta}$ is the descriptive global person-fit statistic; G^2_{β} denotes the test whether $\hat{\beta}$ is significantly greater than 0; G^2_{γ} (sp – HS) denotes the test for detection of spuriously high X_+ scores; G^2_{γ} (sp – LS) denotes the test for detection of spuriously low X_+ scores; and *Joint* denotes that at least one of G^2_{β}, G^2_{γ} (sp – HS), and G^2_{γ} (sp – LS) was significant.

In this study, it was argued that U-shaped PRFs most likely point at spuriously high X_+ scores, and bell-shaped PRFs at spuriously low X_+ scores. Sometimes other explanations for such PRF shapes are possible. For example, suppose an examinee used a crib sheet for the items of medium difficulty. The resulting PRF may be bell-shaped, whereas X_+ would be spuriously high. This example shows that person-fit results should be interpreted with caution, and that additional information on the respondent is needed before definitive conclusions can be drawn. For example, closer inspection of the item-score vector from the example may reveal unexpectedly many consecutive correct answers. This may point at cheating and subsequent person-fit analysis may use indices sensitive to copying (e.g., Cizek, 1999; Sotaridona & Meijer, 2002; Wollack & Cohen, 1998).

The simulation study showed that the logistic regression PRF tests, $G^2_\gamma(\text{sp} - \text{HS})$ and $G^2_\gamma(\text{sp} - \text{LS})$, were conservative with Type I error rates often close to zero. For considerable levels of misfit and varying Type I error rates, 62% through 99% of the item-score vectors with spuriously high X_+-scores (cheating) were detected, and 22% through 85% of the item score vectors with spuriously low X_+-scores (test anxiety) were detected. Fixing the Type I error at five percent led to higher detection rates. However, the relative power of the different statistics for a fixed Type I error rate was approximately the same as that for varying Type I error rates. From a statistical point of view, low Type I error rates may be considered to be an important shortcoming of person-fit methods. However, from a practical point of view, using a conservative person-fit test can be useful, provided that the person-fit test still has power to detect misfit. First, the item-score vectors that are flagged by a person-fit test probably are the most serious cases of misfit. Second, in practice usually a small proportion of respondents exhibit aberrant behavior. Suppose, that 5 percent of the item-score vectors are flagged as misfitting. Based on a conservative person-fit test, we may have confidence that the majority of these cases are the result of aberrant response behavior. This justifies additional analysis of this 5 percent misfitting item-score vectors.

Emons et al. (2002) proposed a comprehensive person-fit methodology in which global, local, and graphical person-fit analysis were integrated. The rationale behind their methodology is that the combination of these person-fit methods may lead to a more accurate decision about misfit or fit. The person-fit methods presented in this study may be part of this person-fit methodology. In particular, logistic regression extends the investigation of local fit, and provides a statistical framework for analyzing the PRFs of large numbers of respondents.

In this study, each PRF was investigated separately. Future research may be aimed at together analyzing the PRFs of a group. The functional data analysis approach (Ramsay & Silverman, 1997) considers each PRF as a functional datum. Thus, for N persons the data consist of N functional observations. Functional data analysis of PRFs may be useful, for example, to determine psychometric properties of a test or a questionnaire, or to investigate cognitive models. In addition, person-fit methods, such as PRF analysis, might be useful to detect outliers in a functional data analysis.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1984). *Revisie Amsterdamse kinder intelligentie test* (*Revision of Amsterdam child intelligence test*). Lisse: Swets & Zeitlinger.

Bowman, A. W. & Azzalini, A. (1997). *Applied smoothing techniques for data analysis*: *The kernel approach with S-Plus illustrations.* Oxford: Clarendon Press.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Erlbaum.

De Koning, E., Sijtsma, K., & Hamers, J. H. M. (2002). Comparison of four IRT models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement, 26*, 302-320.

Douglas, J. & Cohen, A. (2001). Nonparametric function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*, 234-243.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Emons, W. H. M. (2003). Investigating the local fit of item-score vectors. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 289-296). Tokyo: Springer.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the *U3* person-fit statistic. *Applied Psychological Measurement, 26*, 88-108.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2002). *Global, local, and graphical person-fit analysis using person response functions*. Manuscript submitted for publication.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks, CA: Sage.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383-392.

Habing, B. (2001). Nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement, 25*, 221-233.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items.* Mahwah, NJ: Erlbaum.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory*: *Principles and applications.* Boston, MA: Kluwer-Nijhoff.

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*, 133-146.

Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit tests.* New York: Springer Verlag.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331-347.

Junker, B. W. & Sijtsma, K. (2001). Nonparametric item reponse theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211-220.

Klauer, K. C. (1991). An exact and optimal standardized person fit test for assessing consistency with the Rasch model. *Psychometrika, 56*, 213-228.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31*, 19-26.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society (B), 42*, 109-142.

Meijer, R. R. (1994a). *Nonparametric person fit analysis.* Unpublished doctoral dissertation, Vrije Universiteit, Amsterdam, The Netherlands.

Meijer, R. R. (1994b). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311-314.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21*, 99-113.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using IRT based person-fit statistics. *Psychological Methods, 8*, 72-87.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111-120.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* New York/Berlin: DeGruyter.

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430.

Molenaar, I. W. & Sijtsma, K. (2000). *MSP5 for windows. User's manual.* Groningen, The Netherlands: ProGAMMA.

Nering, M. L. & Meijer, R. R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22*, 53-69.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.

Ramsay, J. O. (2000). *TestGraf. A program for the graphical analysis of multiple choice test and questionnaire data* [Computer program]. Montreal, Canada: Department of Psychology, McGill University.

Ramsay, J. O. & Silverman, B. W. (1997). *Functional data analysis.* New York: Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*, 543-568.

Rosenbaum, P. R. (1987a). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40, 157-168.*

Rosenbaum, P. R. (1987b). Comparing item characteristic curves. *Psychometrika, 52,* 217-233.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*, 3-31.

Sijtsma, K. & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79-105.

Sijtsma, K. & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191-208.

Sijtsma, K. & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Simonoff, J. S. (1996). *Smoothing methods in statistics.* New York: Springer-Verlag.

Sotaridona, L. S. & Meijer, R. R. (2002). Statistical properties of the *K*-index for detecting answer copying. *Journal of Educational Measurement, 39*, 115-132.

Strandmark, N. L. & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement, 11,* 355-370.

Trabin, T. E. & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 83-108). New York: Academic Press.

Van den Brink, W. P. (1977). Het verken-effect [The scouting effect]. *Tijdschrift voor onderwijsresearch, 2*, 253-261.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties (Comparability of individual test performance).* Lisse: Swets & Zeitlinger.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267-298.

Vermunt, J. (1997). *ℓEM. A general program for the analysis of categorical data* [Computer program]. Tilburg, The Netherlands: Tilburg University.

Verweij, A. C., Sijtsma, K., & Koops, W. (1996). A Mokken scale for transitive reasoning suited for longitudinal research. *International Journal of Behavioral Development, 19,* 219-238.

Wollack, J. A. & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 21*, 307-320.

Accepted September, 2003.